Update/Le point

Use of time-series analysis in infectious disease surveillance

R. Allard[1]

This article reviews the practical aspects of the use of ARIMA (autoregressive, integrated, moving average) modelling of time series as applied to the surveillance of reportable infectious diseases, with special reference to the widely available SSS1 package, produced by the Centers for Disease Control and Prevention.

The main steps required by ARIMA modelling are the selection of the time series, transformations of the series, model selection, parameter estimation, forecasting, and updating of the forecasts. The difficulties most likely to be encountered at each step are described and possible solutions are offered. Examples of successful and unsuccessful modelling are presented and discussed. Other methods, such as INAR modelling or Markov chain analysis, which can be applied to situations where ARIMA modelling fails are also dealt with, but they are less practical.

ARIMA modelling can be carried out by adequately trained nonspecialists working for local agencies. Its usefulness resides mostly in providing an estimate of the variability to be expected among future observations. This knowledge is helpful in deciding whether or not an unusual situation, possibly an outbreak, is developing.

Introduction

Early identification of an outbreak of a reportable disease is the first step toward an effective intervention to contain it. However, outbreaks are often well under way before public health authorities become aware of them. Time-series analysis based on the Box–Jenkins or ARIMA (*a*uto*r*egressive, *i*ntegrated, *m*oving *a*verage) method models reported cases over time and thereby permits forecasts to be made of expected numbers of reported cases and provides confidence intervals around these forecasts. Having forecasts at hand to compare with the observed numbers of cases can facilitate making a decision as to whether an apparent excess represents an outbreak rather than a random variation. This article discusses the use by nonspecialists of this method of time-series analysis for infectious disease surveillance, paying particular attention to the circumstances that favour or hinder its usefulness.

Many textbooks and other sources (e.g. *1* and *2*) explain the theory of time-series analysis and several statistical packages are available to carry out the calculations involved, e.g. BMDP, S-PLUS and SYSTAT. The reader interested in carrying out ARIMA modelling needs to have access to such a package. Since the Statistical Surveillance System 1 (SSS1) (*3*) is the most user-friendly software for ARIMA modelling, we will refer to it repeatedly in this article. It is produced by the Centers for Disease Control and Prevention (CDC) and is available free on the Internet at http://www.cdc.gov/epo/epi/software.htm. All the figures in this article were generated using SSS1.

ARIMA modelling is theoretically sound and practical, and it is not necessary to have a complete understanding of the underlying statistical theory to apply this method successfully (*2*). However, in order to gain some understanding of the method and be able to apply it prudently, a basic understanding of algebra (square root, reciprocal, logarithm), statistics (mean, moving average, variance, normal distribution, significance, confidence intervals, goodness-of-fit, correlation, and partial correlation)

[1] Senior Epidemiologist, Epidemiologic Surveillance Bureau, Department of Public Health, Montreal General Hospital; and Adjunct Professor, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada. Requests for reprints should be sent to Dr Allard at the following address: Department of Public Health, Montreal General Hospital, 1616 boul. René-Lévesque Ouest, Suite 300, Montreal, Canada, H3H 1P8.

Reprint No. **5868**

and, less importantly, estimation techniques (least-squares method, iterations, convergence), is needed.

All the examples are drawn from experience gained in Montreal, Canada, whose population is 1.7 million, with about 8000 infectious disease notifications per year.

Selection of the time series

Time-series analysis requires a series of observations, repeated at equal time intervals (usually), on the same population. For the forecast intervals to be accurate, the observations should have a normal distribution, with a mean and variance that remain constant over time (a property called "stationarity") (*1*).

The series must not be subdivided into such short intervals that the numbers of observations per interval (case reports, in our context) are so small as to be nonnormally distributed. The interval concerned can be a day, week, a 28-day period, a month, etc. (all series in this article consist of 28-day intervals). However, for rare diseases only a few cases may occur per interval, even if the interval is made long, and ARIMA modelling may therefore not be useful for such diseases.

The longer the series, the better; however, the series should not extend so far into the past as to include periods during which a different case definition was applied or in which any other reporting artifact resulted in a mean number of cases per interval that differs from the mean of recent intervals.

Experience shows that series with a clear periodicity (i.e., the numbers of observations per interval increase and decrease in cycles, generally of 1 year) are more likely to lead to useful forecasts than series without periodicity. Periodicity over 1 year is called seasonality, and for seasonality to become apparent the series should cover at least 2 years.

Should an outbreak have occurred during the period covered by the series, one ought to consider excising from the series the intervals containing the excess cases, taking care to remove a whole year (or several years) of intervals in order to retain any seasonality present in the data.

Since the purpose of ARIMA modelling is to generate useful forecasts, the series should contain data of exactly the same nature as the forecasts required. For example, at the end of each 28-day period, we report provisional numbers of reported cases and these numbers are subsequently updated when delayed reports are received or for many other reasons. For purposes of early outbreak detection, the forecasts are compared to provisional numbers as soon as these become available; thus, the forecasts and the series that generates them should also consist of provisional numbers.

Transformations of the series

As discussed above, for adequate ARIMA modelling a time series should be stationary with respect to mean and variance. If the mean increases or decreases over time, or if the variance does (as indicated by the excursions around the mean becoming smaller or larger over time), the series may need to be transformed to make it stationary, before being modelled.

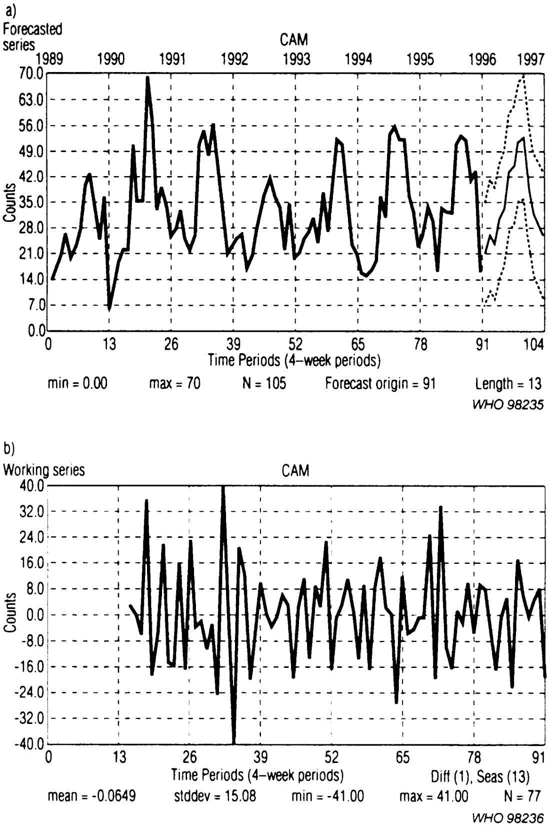
To stabilize the variance, various transformations are available. Each observation can be replaced by, for example, its reciprocal, logarithm or square root. The SSS1 software can analyse the variance of the series and propose a transformation. Our experience indicates that such proposals should not be accepted blindly. The logarithmic transformation, which seems to be the one most often recommended by SSS1, can be applied without fear of complications; however, other transformations, if recommended, should be used with the proviso that the original series be used or another transformation tried if a satisfactory model is not obtained. The reciprocal transformation, in particular, can make it difficult to fit a good model (see below).

If necessary to stabilize the mean the series can be "differenced". In the presence of a secular trend in the series, *regular* differencing is indicated: each observation is replaced by the difference between it and the previous observation. In the presence of clear seasonal variations, *seasonal* differencing is indicated: each observation is replaced by the difference between it and the observation a year before. Both types of differencing can be carried out on the same series.

A simple inspection of the graph of the untransformed series is the most useful approach. If a clear secular trend or seasonal effect is observed, the series should be differenced, otherwise not. Similarly, only if the variance clearly increases or decreases should some transformation be applied.

In Fig. 1a) the solid line shows a series with a clear yearly periodicity and slight fluctuations in its yearly mean. Fig. 1b) shows the same series after seasonal and regular differencing. The first year of observations are lost because they cannot be seasonally differenced. The differenced series appears stationary, with a variance that seems to decrease only slightly over time. No transformation was required, in our opinion.

Fig. 1. a) **Provisional numbers of notifications of *Campylobacter* infections, per 28-day interval, Montreal, Canada, 1989–95, followed by forecasts for 1996, based on the differenced series** (bold line: observations; thin line: forecasts; dashed lines: 95% forecast intervals). b) **Same series after regular and seasonal differencing.**



Autocorrelation analysis and model specification

Once the series has been made stationary, the next step is to analyse its autocorrelation structure. Autocorrelations are the correlations between each observation and the previous one (lag one) or the previous observation but one (lag two) and so on. Partial autocorrelations are the same, except that the effect of the intervening observation(s) is removed. Clearly, for lag one, since there is no intervening observation, the autocorrelation and the partial autocorrelation are the same. SSS1 and other software packages present autocorrelations of both types graphically, up to a higher lag than is usually needed, and show which autocorrelations are significantly different from zero.

The size of the autocorrelations and partial autocorrelations of various lags guides the selection of terms to include in the initial ARIMA model. To do this optimally requires a sophisticated knowledge of the theory behind ARIMA modelling; however, in most instances a few simple principles are enough to do the job adequately.

An ARIMA model can include two sets of terms (or parameters): "autoregressive" (AR) terms and "moving average" (MA) terms. Each set can include terms of any lag, termed "degree" in this context. Autoregressive terms relate the *observation* made at time *t* in the series to the observation made at time *t*–1 (first degree), or to the observation made at *t*–2 (second degree) and so on. Moving average terms relate the *error* (difference between observation and estimated value) at time *t* to the error at times *t*–1, *t*–2, etc. Both sets can also include seasonal terms (of degree 12, 13, 52 etc. depending on the interval between observations) and their multiples.

The initial model should include moving average terms that have the same number of degrees as the significant autocorrelations and autoregressive terms that have the same number of degrees as the significant partial autocorrelations. It may also require a constant term, especially if a series showing a time trend has been left undifferenced. SSS1 offers autoregressive and moving average terms, each type up to degree five, plus two seasonal terms (1 or 2 years); it also warns the user about including a constant term, when one is indicated.

ARIMA modelling is based only on the mathematical properties of the series and not on the dynamics of infectious disease transmission. The nature of the observed events is irrelevant.

Estimation

Once the initial model has been specified, estimates can be made for the parameters, i.e. numerical values for the parameters can be derived from the observations in the series. This does not usually require any decision on the part of the user. Estimation is iterative, by the least squares method. Should the estimation procedure fail to converge, the user may choose initial values different from the default ones, but we have never had to do this.

Model refinement

Apart from the estimated parameter values, the procedures used by most software generates confidence intervals, significance levels, model sums of squares and several closeness-of-fit statistics. The software can also display graphs of the observed and the estimated values for the series and of the corresponding residuals (see below). How best to use these indicators to select which terms to retain in the model and which to exclude is arcane. The meaning of each individual indicator is fairly clear, but how to combine them to arrive at a decision is far less so. In practice, we have found the significance level of the parameters to be the most useful in this respect. Excluded first from the model are the least significant parameters, the remaining parameters are reestimated (together with their significance levels), the least significant are excluded from the parameters remaining, and so on reiteratively until only the most significant parameters remain. As far as the significance level is concerned, a permissive rather than a restrictive approach is indicated, as has been recommended for other forms of modelling (*4*); in this way parameters with *P*-values less than, say 0.1, can be retained rather than using a stricter cut-off level of $P = 0.05$ or $P = 0.01$.

In this respect, the following points should be borne in mind. 1) Little or no harm is done by leaving nonsignificant parameters in the model; the confidence intervals of the other parameters may be unnecessarily widened, but this effect seems to be small in ARIMA modelling (e.g. compared with logistic regression). Parsimony is not as important when the objective is useful forecasting as when it is the theoretical understanding of the epidemic process, which ARIMA models are not designed to clarify. 2) When more than one parameter is nonsignificant, the higher-degree parameters should be eliminated first, and the model tested again. 3) There should be greater reluctance to drop seasonal parameters than regular parameters, because of their importance for forecasting.

The residuals, i.e. the difference between each past observation and its expected value according to the model, should also be inspected; values for these are provided by the statistical package. Residuals should ideally be small, as frequently positive as negative, and show no secular or seasonal trend.

Forecasting

Once a satisfactory model has been obtained, it can be used to forecast expected numbers of cases for a given number of future time intervals. A starting point and a duration have to be chosen, with the last observation in the series being a natural starting point for forecasting. The number of intervals into the future that a forecast should be attempted depends on the degree of the model (i.e., the highest degree of any parameter it contains) and care should be taken in making forecasts for a number of intervals greater than the degree of the model. If the model is of the first degree, each forecasted value depends only on the previous value in the series; for the first forecasted value this would usually be the last observation, but the second forecasted value would be based entirely on the first forecast, and so on. Basing subsequent forecasts on previous forecasts so greatly increases their confidence intervals as to make them rapidly useless. In contrast, a seasonal term makes it more reasonable to try and forecast for a whole year, because the later forecasts are based at least in part on observations (because of the seasonal parameter(s)) as well as on previous forecasts (because of the lower-degree parameters).

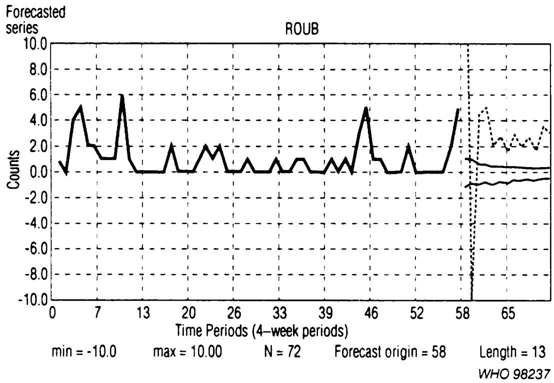
A simple visual inspection of the forecasted series following the observed series can make it easier to decide whether the forecasts make sense in relation to the series of observations. If there is a clear time trend in the original (undifferenced) series, it should appear in the forecasts. A clear periodicity in the series should also be reflected in the forecasts. Forecasts that appear unreasonable to the experienced eye should prompt a reassessment of the whole process.

It is easy to test the model by choosing as the starting point for forecasing not the last observation but an earlier one since subsequent past observations can then be compared with their "forecasted" value; however, since such observations have served to estimate the model parameters, this cannot be considered a rigorous test.

Fig. 1 a) shows a series of *Campylobacter* notifications, followed by forecasts for 1996 generated by a pure moving average model of degrees 1, 2, 13 and 26. These forecasts are credible: they reflect the seasonal variations and the confidence limits are plausible.

Fig. 2 shows a series of measles notifications. A reciprocal transformation was applied, as recommended by SSS1; the model is both autoregressive, of degrees 1 and 2, and moving average, of degree 13. The forecasts are patently absurd since the estimated numbers of notifications sometimes lie outside their own confidence interval. Fig. 3 shows forecasts generated from the same series, but untransformed. These results are more plausible, but their usefulness is limited to showing that more than two or three notifications per interval probably represents an excess. Fig. 3 also shows that the lower confidence limit and the forecasts themselves can be negative. Negative forecasted values are taken to be zero, but they throw doubt on the adequacy of the model. In this series, there is a preponderance of small values (0

Fig. 2. **Provisional numbers of notifications of measles infections, per 28-day interval, Montreal, Canada, 1990 to mid-1994, followed by forecasts based on the differenced and reciprocally transformed series, for the next 13 intervals.**



and 1), which does not meet the normality assumption; hence the poor performance of the method.

Using the forecasts

Series should be identified precisely; for example, "Acute hepatitis B, provisional numbers of confirmed cases, per month". Otherwise, the forecasts may later be compared with the wrong series of observations.

Fig. 3. **Provisional numbers of notifications of measles infections, per 28-day interval, Montreal, Canada, 1990 to mid-1994, followed by forecasts** (dotted line) **based on the original series, for the next 13 intervals** (dashed line: 95% forecast intervals).

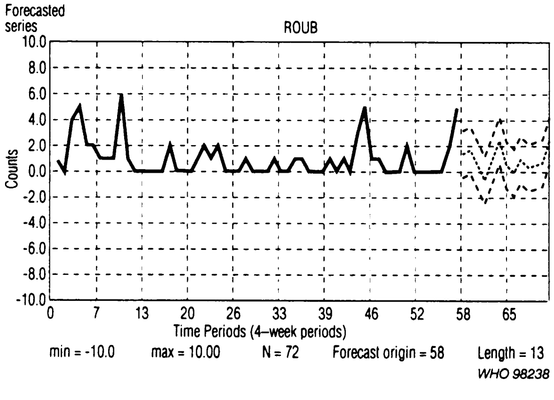
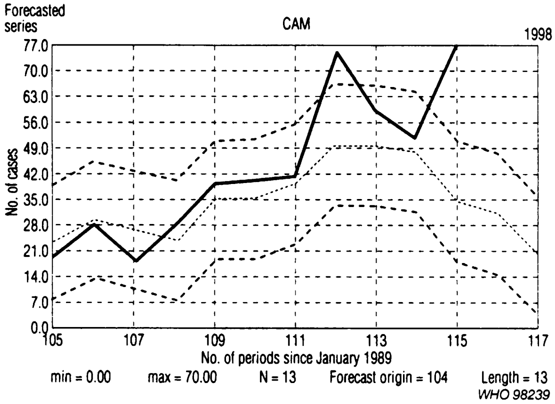


Fig. 4. **Forecasted and observed numbers of notifications of *Campylobacter* infections, per 28-day period, Montreal, Canada, 1997** (solid line: observations; dotted line: forecasts; dashed lines: 95% forecast intervals).



The forecast intervals are as important as the forecasts themselves. Clearly, the upper limit is of particular interest, since it suggests that these is a significant excess of reported cases if it is exceeded by the observations. The lower limit can sometimes be of use if the implementation of a disease control programme is accompanied or followed by a significant decrease in reported cases.

In practice, the most important point is to keep the forecasts at hand, ideally on display, and to write in new observations as soon as they become available. Otherwise, the exercise will have no chance of helping to detect outbreaks sooner than otherwise possible.

Fig. 4 presents an example of the use of forecasts for the routine surveillance of infectious diseases. Shown are the forecasted numbers of *Campylobacter* notifications for the 13 four-week periods of 1997, with their 95% confidence intervals. The forecasts are based on the series presented in Fig. 1, extended through 1996, and are similar to the forecasts for 1996. Significant excess numbers of cases occured during the 8th and 11th periods of 1997, which are not related to known outbreaks of the disease. Demonstrating that the first excess was outside the forecasted range was an incentive to investigate the reported cases, which we do not do routinely; unfortunately, however, this did not uncover a common source or mode of transmission. Two periods later, the occurrence of an even greater excess at a time when the number of notifications should have decreased prompted us to intensify the investigation, which is continuing.

Updating the forecasts

How often the forecasts need to be updated depends on the degree of the model and on the length of the series of forecasts. A model without seasonal terms will need to be updated several times a year. Confidence intervals that widen rapidly as time increases from the starting point of the forecasts also indicate a model that needs frequent updating. In contrast, a model with at least one strong seasonal parameter may yield good forecasts for a year.

Should an excess seem to occur which would, if confirmed as an outbreak, require a large-scale public health intervention, it may be prudent to repeat the modelling and forecasting with a series that goes right up to the point where the apparent excess begins.

The updating can be done in two ways. The model may be saved and can be reapplied to the original series with extra observations added at the end to give forecasts based on a later starting point. Alternatively, a new model can be fitted to the longer series; this is probably preferable, since fitting a model is quick, especially when the old model is used as a guide, and it makes better use of the additional observations.

Discussion

ARIMA modelling is a useful tool for interpreting surveillance data and has helped us interpret more rapidly any increases in some common reportable diseases. However, we have not been able to model adequately some diseases that are rare but important because of their severity, such as meningococcal meningitis.

Other modelling techniques have been developed for such situations. The INAR (*in*teger-valued *a*uto*r*egressive) model (*5*) does not require that the observations have a normal distribution, and it generates confidence intervals that are integers, as are the observations themselves. However, the currently available software for INAR is not as flexible and user-friendly as SSS1 (Lambert J, Ranger N, Roy R. INAR, XINAR and INARG programs, mimeographed text, and diskette, 1993) and we have not so far been able to apply INAR usefully.

Use of Markov chain analysis (*6*) can provide the probability that the next observation will be 0, 1, etc. up to the largest observation in the series, which can be useful for forecasting very rare diseases. However, there is no widely available statistical package to perform the calculations. Such calculations are straightforward if the range of observations is not too large and if probabilities are required only for the next interval. To obtain probabilities for several consecutive intervals, matrix multiplication is required. Markov chain analysis can be applied to the series of measles notifications shown in Fig. 3 by creating a two-by-two table of how often each observed value (here, 0 to 6) is followed by itself or each other value, and transforming the frequencies of these pairs into probabilities by dividing them by the overall frequency of the first value of the pair. This gives an estimate of how likely it is that a 0, say, will be followed the next period by a 0, a 1, a 2 etc. In this analysis, we collapsed the rows and columns corresponding to 3, 4, 5 and 6 cases because of small numbers. This gives the probabilities of observing 0 case one period, followed by 0, 1, 2 or $\geqslant 3$ cases the next period, as 0.63, 0.20, 0.10 and 0.07, respectively, based on 30 pairs of observations. The probabilities of observing one case one period, followed by 0, 1, 2 or $\geqslant 3$ cases the next period are 0.53, 0.27, 0.13 and 0.07, respectively, based on 15 pairs of observations. Thus, if 0 cases or 1 case is observed in a period, which is the most common occurrence, then it takes at least 3 cases the next period for the sequence to be an unlikely occurrence ($P = 0.07$). This is almost the same as the upper limit given by ARIMA modelling. This example shows that in some situations a useful Markov chain analysis can be carried out with no more than pencil and paper.

Conclusion

In our experience, the usefulness of forecasting expected numbers of infectious disease reports consists not so much in detecting outbreaks or providing probability statements, but in giving decision-makers a clearer idea of the variability to be expected among future observations. This becomes one more element in the subjective determination of whether an unusual situation is or is not developing. In our experience, variability has often been much larger than we would have otherwise expected, and this has helped us decide whether further surveillance and/or a public health intervention was required.

Ideally, every reported case of a transmissible disease should be investigated and the causes of its occurrence fully understood. Nowhere is this possible, for scientific and practical reasons. Since the smaller the population on which observations are made, the larger their variability is likely to be, we believe the intelligent application of ARIMA modelling in smaller jurisdictions can help focus public health efforts on unusual situations, and avoid investigating random fluctuations — an effort unlikely to be profitable.

Acknowledgements

Dr J. Lambert in thanked for providing software to carry out INAR modelling; Ms M. Cardinal for providing information on Markov chain analysis and for her comments; and Ms L. Bédard and Dr A. Dean for their comments.

Résumé

Application de l'analyse des séries chronologiques à la surveillance des maladies infectieuses

L'article envisage les aspects pratiques de la méthode ARIMA (*autoregressive, integrated, moving average*) appliquée à la surveillance des maladies infectieuses déclarables, et notamment le système SSS1 (*Statistical Surveillance System*) élaboré par les *Centers for Disease Control and Prevention.*

Les étapes principales nécessaires de la modélisation ARIMA sont les suivantes : choix des séries chronologiques, transformation des séries, choix du modèle, estimation des paramètres, prévisions et mise à jour des prévisions. Les séries choisies pour le modèle doivent couvrir au moins 2 ans pour, le cas échéant, pouvoir rendre compte des variations saisonnières de fréquence de la maladie. En cas de flambée épidémique, la période correspondante peut être exclue du modèle de façon à refléter la situation habituelle. La série doit être découpée en périodes (jours, semaines, mois, etc.) suffisamment longues pour que le nombre d'observations (de cas) ait une distribution normale dans chacune d'elles. Il est parfois nécessaire de transformer la série, pour faire en sorte que sa moyenne et sa variance soient «stationnaires», c'est à dire restent constantes au cours du temps, condition d'existence du modèle. Un modèle ARIMA peut inclure des termes qui traduisent l'influence sur chaque période de la période antérieure, de celle qui précède cette période antérieure, et ainsi de suite, ou encore des périodes correspondantes des années précédentes (termes de saisonnalité). Le choix des termes à inclure dans le modèle repose sur la corrélation entre le nombre de cas dans la période et chacune des périodes précédentes. La longueur de la période pendant laquelle on peut faire des prévisions fiables dépend du modèle, et en particulier de la présence ou non de termes de saisonnalité dans le modèle. La fréquence avec laquelle les prévisions doivent être remises à jour dépend de même du modèle, mais aussi de l'intervalle qui sépare les futures observations des prévisions correspondantes. Plus elles sont proches, plus durable est la validité du modèle.

On trouvera décrits et discutés des exemples de modèles plus ou moins adéquats. Il existe d'autres méthodes, le modèle INAR et l'analyse des chaînes de Markov par exemple, applicables en cas d'échec de la méthode ARIMA; elles sont toutefois moins pratiques.

La modélisation ARIMA peut être mise en œuvre par des non-spécialistes convenablement formés travaillant pour des organismes locaux, et permet d'affirmer que l'apparition d'un grand nombre de cas pourrait représenter le début d'une épidémie. Il ressort toutefois de notre expérience que son intérêt principal est de donner une estimation de la variabilité probable des futures observations. Il est très souvent arrivé que la variabilité soit beaucoup plus importante que prévu, et nous avons pu décider de la nécessité d'une intervention.

References

1. **Abraham B, Ledolter J.** *Statistical methods for forecasting.* New York, John Wiley & Sons, 1983: 281–321.
2. **Wilkinson L, Hill MA.** *Systat for DOS, Version 6. Advanced applications.* Evanston, IL, 1994: 589–621.
3. **Haddad S, Basha M, Rapose W.** *Statistical software for public health surveillance.* A.P.C. Systems, Inc. and Centers for Disease Control, Washington, DC: 22–82.
4. **Greenland S.** Modeling and variable selection in epidemiologic analysis. *American journal of public health,* 1989, **79**: 340–349.
5. **Du JG, Li Y.** The integer-valued autoregressive (INAR(p)) model. *Journal of time series analysis,* 1991, **12**: 129–142.
6. **Hillier FS, Lieberman GJ.** *Introduction to operations research.* San Francisco, CA, Holden-Day, 1967: 402 ff.